View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Dokumenten-Publikationsserver der Humboldt-Universität zu Berlin

R. Harald Baayen, Antti Arppe
University of Alberta, University of Helsinki
baayen@ualberta.ca, antti.arppe@helsinki.fi

Statistical classification and principles of human learning

In the application of any statistical analysis method to the modeling of linguistic phenomena, a recurring question is how to understand the statistical results from a cognitive perspective. Although quantitative models may provide detailed and useful insights into which factors enhance the probability of particular linguistic phenomena, they tend leave unanswered how actual speakers come to learn and use their language in the way they do.

The present study addresses this question by introducing a new, parameter-free model for linguistic choice behavior based on naive discriminative learning that is driven fully and only by the distributional properties of its input. The learning principles on which this model is based, the so-called Rescorla-Wagner equations (see appendix), were first proposed by Wagner and Rescorla in 1972 (Wagner & Rescorla 1972), and have proved to be amazingly fruitful in psychology as a model for human and animal learning (Miller, Barnet, & Grahame 1995). A technical innovation due to Danks (2003) makes it possible to estimate the weights of the Rescorla-Wagner equations when learning has reached a state of equilibrium. Baayen, Milin, Filipović Durdević, Hendrix, and Marelli (submitted) incorporated the equilibrium equations of Danks (2003) into a general discriminative learning model that is naive in the sense of naive Bayes classifiers. These authors show that naive discriminative learning provides accurate predictions of response latencies in the visual lexical decision task. The model reproduces a wide range of effects in the morphological processing literature with a minimum of representational assumptions, using a learning engine that, in its simplest form, has no free parameters.

In this paper, we pit this parameter-free statistical engine derived from human learning principles against several well-established statistical classifiers: random forests (Breiman 2001; Strobl, Malley, & Tutz 2009), support vector machines (Vapnik 1995), memory-based learning (Daelemans & Bosch 2005) and polytomous logistic regression (according to the one-vs.-rest heuristic, see e.g. Arppe (2008)).

As our linguistic example case, we have selected the near-synonymous set of the four most frequent Finnish verbs denoting THINK, namely *ajatella*, *miettiä*, *pohtia*, *harkita* ‘think, reflect, ponder, consider’, which have been comprehensively studied by Arppe (2008) using newspaper and Internet newsgroup discussion corpora. Altogether 3,404 occurrences of these four THINK verbs and their sentential contexts were analyzed in terms of their morphological and lexical as well as syntactic structure (following Functional Dependency Grammar, (Tapanainen & Järvinen 1997)), supplemented with semantic and structural subclassifications. Of some 6000 contextual features, 46 were selected for the present study, as these 46 emerged from previous analyses as the most predictive ones when taken together. This subset of predictors included the most common morphological properties and general semantic characteristics of the verb-chain in which the think verb occurred, and detailed information on the syntactic structure (functional roles and various subclassifications) linked with the think verbs in their sentential context. Arppe (2008) observed that using polytomous logistic regression (with any of several common heuristics) as a classifier seems to reach a ceiling at a Recall rate of roughly two-thirds of the sentences in the research corpus. The

results could not be substantially improved with the addition of further granularity in semantic and structural subclassification of the syntactic roles, and effectively similar results were obtained when partially varying (even randomly) the selection of contextual features.

TABLE 1. CLASSIFICATION DIAGNOSTICS FOR FIVE MODELS FITTED TO THE FINNISH DATA SET (N = 3404).

	$\lambda_{prediction}$	$\tau_{classification}$	Recall (proportion correct)
polytomous logistic regression	0.368	0.488	0.645
support vector machine	0.334	0.461	0.626
memory-based learning	0.286	0.422	0.599
random forests	0.326	0.455	0.621
naive discriminative learning	0.349	0.473	0.634

The classification results for the four statistical models and the naive discriminative learning model are summarized in Table 1. The measure for proportionate reduction of prediction error, $\lambda_{prediction}$, tells us how much better the models perform by using the selected set of predictors compared to what would be achieved by systematically selecting the most frequent verb in the data, while the measure for proportionate reduction of classification error, $\tau_{classification}$, informs us how much better the models reproduce, in the long run, the actually occurring proportions of verbs evident in the data, in comparison to the baseline case of homogeneous proportionate distribution (Menard 1995). As the results for random forests may change slightly across different runs, also a mean Recall =0.622 was estimated for a series of 50 random forests (range: 0.617-0.626). From Table 1 we learn that polytomous regression performs best, followed by naive discriminative learning. A proportions test comparing the top two models suggests they have equivalent recall. What we can conclude at this point is that discriminative learning performs as well as other established classifiers, at least on this data set. Interestingly, naive discriminative reading achieves this level of accuracy without a single free parameter, and therefore provides the theoretically most parsimonious fit of all models surveyed here.

At a high level of abstraction, each of the five models surveyed above provides a good characterization of a Finnish native speaker's knowledge of the optimal choice of a think verb given morphological, syntactic and other contextual information. Although roughly equivalent in terms of predictive accuracy, it is only memory-based learning and naive discriminative learning that have some cognitive plausibility — we believe it is unlikely that the brain would actually be searching for support vectors, that it would be estimating beta weights, or that it would be constructing forests of conditional inference trees. Memory-based learning is an attractive paradigm for probabilistic inference in language processing, that is in many ways compatible with usage-based and exemplar-based approaches. A potential disadvantage of memory-based learning is that it requires vast amounts of memory for the exemplars, combined with on-line computations on nearest neighbor sets. By contrast, discriminative learning assumes that the adult competence is the result of a long process of discriminative learning. This model is extremely sparse in the number of representations and connections required: for the present data set, all we need is 4 representations , one for each of the think verbs, 46 representations for the binary predictor values, and 4 ∗46 = 148 connection weights. The support for a particular verb given the input is calculated straightforwardly by summation of the weights on the connections linking the input predictors to the verb.

Exemplar knowledge is not stored explicitly in the form of 3404 exemplar vectors, but implicitly in just 148 connection weights. We hypothesize that naive discriminative learning implements the simplest possible mathematical characterization of probabilistic linguistic competence, compatible with the insight that grammar is usage-based, but without assuming that usage is calculated over an entire exemplar space. Instead, we assume that usage is acquired piecemeal in a much simpler weight space.

References

Arppe, A. 2008. *Univariate, Bivariate and Multivariate Methods in Corpus-based Lexicography. A Study of Synonymy*. Helsinki: Department of General Linguistics, University of Helsinki.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P. and Marelli, M. submitted. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning.

Breiman, L. 2001. *Random forests. Machine Learning* 45, 5-32.

Daelemans, W. and Bosch, A. Van den. 2005. Memory-based Language Processing. Cambridge: Cambridge University Press.

Danks, D. 2003. Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology* 47(2), 109-121.

Menard, S. 1995. Applied logistic regression analysis. Thousand Oaks: Sage Publications.

Miller, R. R., Barnet, R. C. and Grahame, N. J. 1995. Assessment of the Rescorla-Wagner model. *Psychological Bulletin* 117(3), 363-386.

Strobl, C., Malley, J. and Tutz, G. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods* 14(4), 26.

Tapanainen, P. and Järvinen, T. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, April 1997, 64-71. Washington, D.C.: Association of Computational Linguistics.

Vapnik, V. 1995. The nature of statistical learning theory. Berlin: Springer-Verlag.

Wagner, A. and Rescorla, R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F. (eds.), *Classical Conditioning II*, 64-99. Appleton-Century-Crofts.

Appendix

Let PRESENT(X, t) denote the presence of a cue (predictor value) or outcome (one of the four Finnish THINK verbs) X at time t, and ABSENT(X, t) denote its absence at time t. The Rescorla-Wagner equations specify the association strength V_i^{t+1} of cue Ci with outcome O at time $t + 1$ using a recurrence equation, as follows:

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{1}$$

The change in association strength ΔV_it defined as

$$\Delta V_i^t = \begin{cases} 0 & \text{if ABSENT}(Ci,t) \\ \alpha_i\beta_1(\lambda - \Sigma_{\text{PRESENT}(Cj,t)} Vj) & \text{if PRESENT(Cj,t) \& PRESENT(O, t)} \\ \alpha_i\beta_2(0 - \Sigma_{\text{PRESENT}(Cj,t)} Vj) & \text{if PRESENT(Cj,t) \& ABSENT(O, t)} \end{cases} \tag{2}$$

The equilibrium equations of Danks (2003),

$$\Pr(O|Ci) - \Pr(Cj|Ci)Vj = 0 \tag{3}$$

make it possible to estimate the weights for an ‘adult’ system by solving the above set of equations using the co-occurrence vector of a specific outcome (verb) given the different predictor values and the co-occurrence matrix of predictor values.